Discovering new RNA targets in Genomic DNA

Stephen R. Holbrook, Inna Dubchak, and Richard J. Carter

Computational & Theoretical Biology Department, Physical Biosciences Division, & National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

In light of recent discoveries of its roles in varied cellular processes, RNA has become an exciting therapeutic target. The identification of potential RNA targets would allow a corresponding development of novel therapeutic approaches. Until recently, there has been no successful computational approach for identification of genes encoding novel functional RNAs (fRNAs) in genomic sequences. We have developed a machine learning approach that contrasts known RNA and non-coding sequences to extract common features that can distinguish functional RNAs. These trained computational machines are then used for prediction of new RNA genes in the unannotated regions of prokaryotic and archaeal gemones.

The *E. coli* genome was used for development, but we have applied this method to several other bacterial and archaeal genomes. Computational neural networks based on nucleotide composition were 80-90% accurate in jackknife testing experiments for bacteria and 90-99% for hyperthermophilic archaea. We also achieved a significant in accuracy by combining these predictions with those obtained using a second set of parameters consisting of known RNA sequence/structure motifs and the calculated free energy of folding. Several known fRNAs not included in the training datasets were identified as well as several hundred predicted novel RNAs. These studies indicate that there are many unidentified RNAs in simple genomes that can be predicted computationally as a precursor to experimental study. We also noted that in several cases functional segments in the untranslated regions of mRNA were correctly identified by our networks.

Preliminary results indicate that the method is applicable to fRNA prediction in higher organisms, including the human genome. The method, which is simple to run and is available via the web (http://rnagene.lbl.gov), may yield the discovery of thousands of novel fRNAs in these higher organisms.